Indian J. Phys. **65B** (5), 467-473 (1991)

Analysis of codon usage : Influence of the purine and pyrimidine combinations in symmetric/asymmetric modes

J Seetharaman and R Srinivasan
Department of Crystallography and Biophysics,
University of Madras, Madras-600 025, India

Abstract : Using the codon usage table from the Gen Bank Genetic Sequence Data for 2186 genes of various individual organisms, 758030 codons are used for the analysis of the pattern of the combination of the purine (R) and pyrimidine (Y) bases in the three positions of the 64 triplets that code for the 20 amino acids. The 64 codons are grouped into 4 types as (1) RRR (3R's), (2) YYY (3Y's), (3) RRY (2R1Y) and (4) YYR (2Y1R). Here the 3rd and 4th types of combinations are considered irrespective of their positions. The available data is split up into three catagories as (1) Global (GLO), (2) Higher (HIG) and (3) Lower (LOW). The order of preference is 3R, 2Y1R, 3Y and 2R1Y (for GLO), 3R, 3Y, 2Y1R, 2R1Y (for HIG) and 2R1Y, 3R, 2Y1R, 3Y (for LOW). The two groups HIG and LOW differ much in their preferences. The analysis from symmetric/asymmetric mode of combinations show that the behavior found in GLO is uniformly reflected in HIG but not in LOW. In all the cases, it is found that RRR is used maximally and YRR is least used.

Keywords : 64 codons, purine-pyrimidine, possible combinations, symmetric/asymmetric mode analysis.

PACS No : 87.10.+e

1. Introduction

The available voluminous protein and nucleic acid sequence data has enabled us to carry out some analysis in codon usage, dinucleotide occurrences, amino acid compositions, Watson Crick base pair ratios and purine-pyrimidine ratios. Though some analysis on purine and pyrimidine ratios, occurrences, preferences and interactions are available (Berger 1978, Clary and Wolstenhome 1985, Fitch 1976, Hanai and Wada 1988, Haran *et al* 1987, Jaroslav and Jan Mrazck 1987, Jukes 1978, Wachters Hauser *et al* 1988), recently we have carried out some analysis of the purine and pyrimidine combinations in symmetric/asymmetric mode of occurrences. The 64 codons arise out of the possible combinations of the purine (R=A, G) and pyrimidine (Y=C, U) in each one of the three positions. Here depending on the bases (either R or Y) in each position, they can be grouped into various categories and their frequency of occurrences can be studied. The type of possible combinations are (1) RRR (3R), (2) YYY (3Y), (3) RRY (2R1Y) and (4) YYR (2Y1R).

These four combinations are considered irrespective of their positions (i.e., for instance in 2R1Y it may be RRY or RYR or YRR). The analysis yields additional results when studied from the point of view of the symmetric/asymmetric mode of combinations. The symmetric mode consists of (1) RRR, (2) YYY, (3) RYR and (4) YRY and similarly the asymmetric mode consists of (1) RRY, (2) YRR, (3) YYR and (4) RYY. Here, we have carried out the analysis in two parts with part one consisting of the combinations of 3R's, 3Y's, 2R1Y's and 2Y1R and the second part consisting the symmetric/asymmetric mode combinations.

2. Materials and method

The codon usage tabulated by Aota *et al* (1988) based on Gen Bank Genetic Sequence Data Bank (Release 50.0 1987) for various individual organisms as well as viruses and organella, is used in the present analysis. For the sake of

Table 1. Various symmetric/asymmetric modes*.

	G	C	U	A
G	RRR	RYR	RYR	RRR
	RRR	RYR	RYR	RRR
	RRY	RYY	RYY	RRY
	RRY	RYY	RYY	RRY
C	YRR	YYR	RRY	YRR
	YRR	YYR	YYR	YRR
	YRY	YYY	YYY	RYR
	YRY	YYY	YYY	YRY
U	YRR	YYR	YYR	YRR
	YRR	YYR	YYR	YRR
	YRY	YYY	YYY	YRY
	YRY	YYY	YYY	YRY
A	RRR	RYR	RYR	RRR
	RRR	RYR	RYR	RRR
	RRY	RYY	RYY	RRY
	RRY	RYY	RYY	RRY

*This table gives the possible combinations of the purine (R) and pyrimidine (Y) in the three positions.

convenience, the entire sequence data is recast into two groups and considered as it is for the global behaviour. The two groups are (1) Higher Organisms (HIG-consists of Hum, Ham, Mus, Rat, Bov, Rab, Chk, Xel, Dro, Mze and Ysc-1248 genes), (2) Lower Organisms (LOW-consists of Eco, Sty, Tip, ADB, Hiv, Hs1, Hs4, Lambda, Pt4, Pt7, Mpocp and Topcp-938 genes) and (3) Global (from Hum to Tobcp-2186 genes) (refer Aota *et al* 1988 for abbreviation). The type of analysis that has been carried out is as follows. According to the occurrence of

a given base R or Y in each one of the position I, II and III of the codons, they are grouped into various possible combination type.

3. Results and discussion

Table 1 gives the possible combinations of the R and Y in all the 64 codons. It is clear from the Table 1 that number of codons in 2R1Y and 2Y1R types are 24 and in the remaining cases it is 8.

The analysis for the global data yields the following results. Of the four groups, 2Y1R is used maximally followed by 2R1Y (Figure 1, the frequencies are plotted in histograms). The reason may be that each is accounted by 24 codons. Among the rest, 3Y is found to be the least used (Table 2). Now in order to

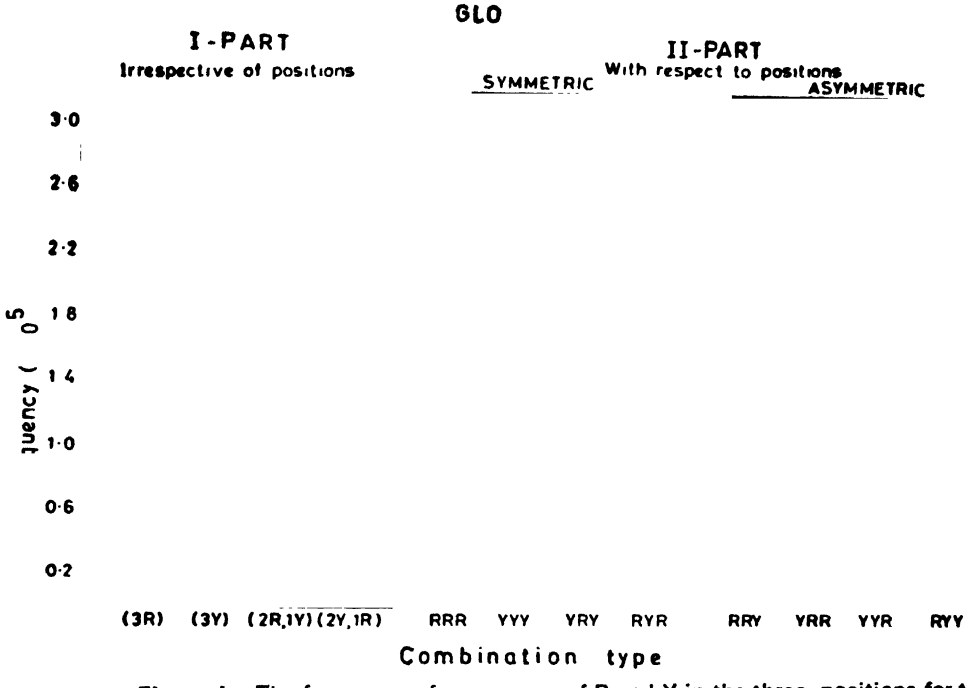

Figure 1. The frequency of occurrence of R and Y in the three positions for the Glo is represented by histograms in I'' part and their occurrences interms of symmetric and symmetric modes are represented in II'' part.

remove the discripancies, the number of codons in each combination type are normalized in terms of the number of codons. Thus in normalized condition, it is found that RRR has the highest usage and the least being that for 2R1Y (Figure 1a). When such an analysis for LOW and HIG are compared, it is clear that they differ widely and in all cases HIG is higher except in 2R1Y case (Figures 2 and 3). When they are normalized in terms of number of codons (Table 3) the order of preference are (1) 3R, 2Y1R, 3Y, 2R1Y (for GLO), (2) 3R, 3Y, 2R1Y and 2R1Y (for HIG) and (1) 2R1Y, 3R, 2Y1R and 3Y (for LOW) (Figures 1a, 2a, 3a).

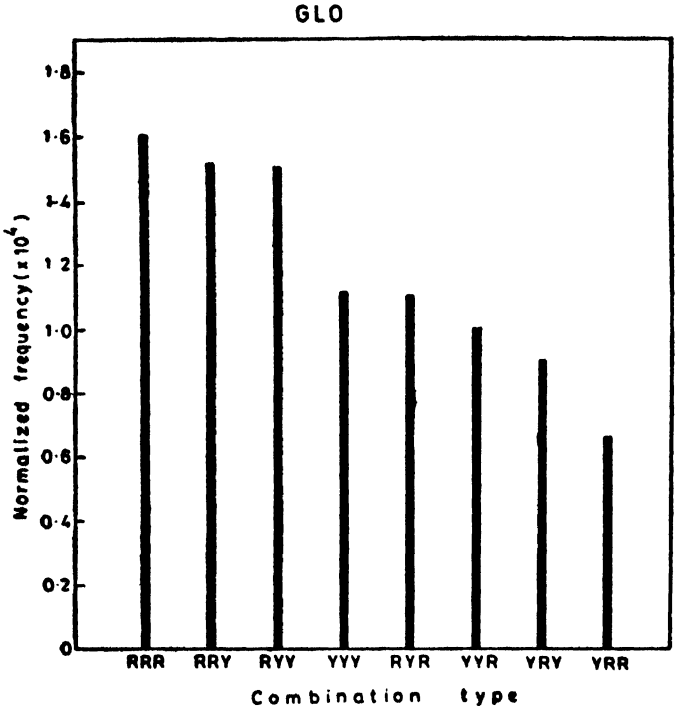

Figure 1a. The normalized frequency of occurrence of all the combination types in GLO.

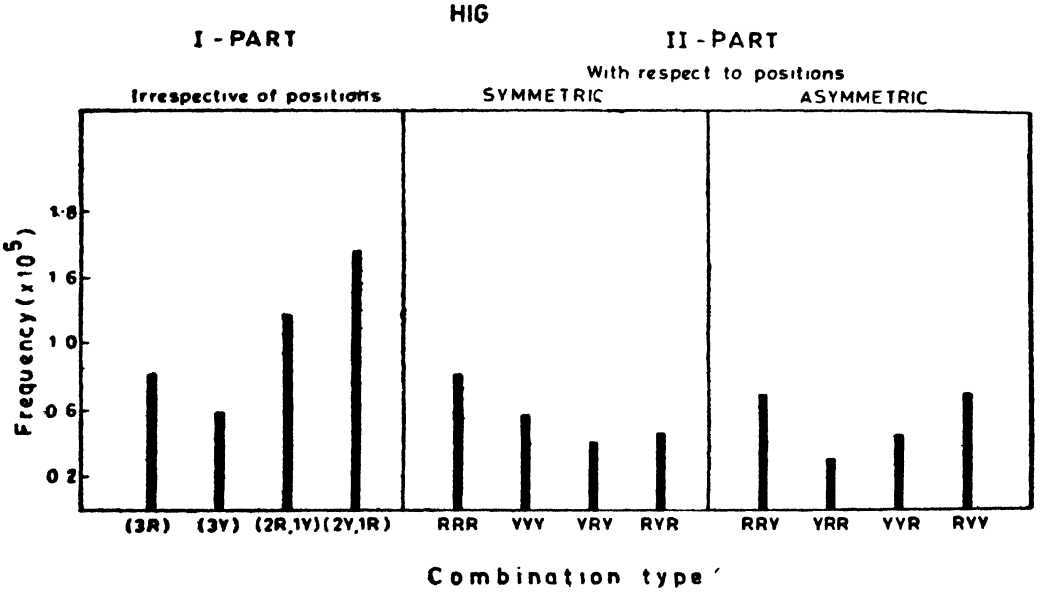

Figure 2. The frequency of occurrence of R and Y in the three positions for the HIG is represented by histograms in I" part and their occurrences of symmetric and symmetric modes are represented in II" part.

From the above results it is clear that the behaviour embodied in GLO is reflected in the groups HIG with respect to most and least preferences and the intermediate

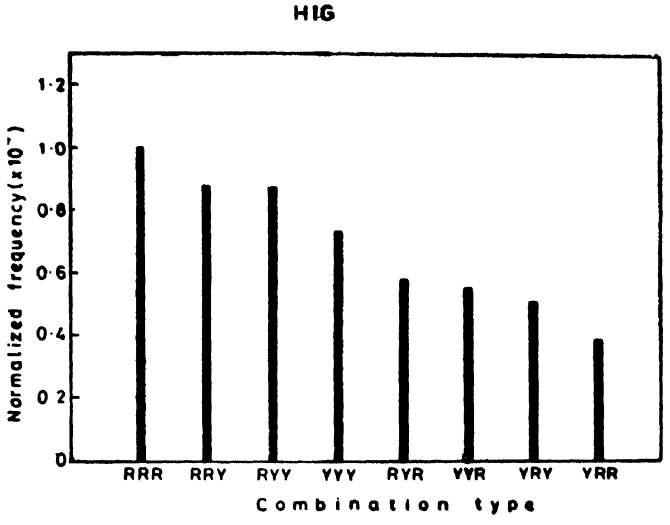

Figure 2a. The normalized frequency of occurrence of all the combination types in HIG.

two modes of combinations 2Y1R and 3R are interchanged. But the LOW groups completely deviate from the GLO and nearly opposite tendency of their usage are found.

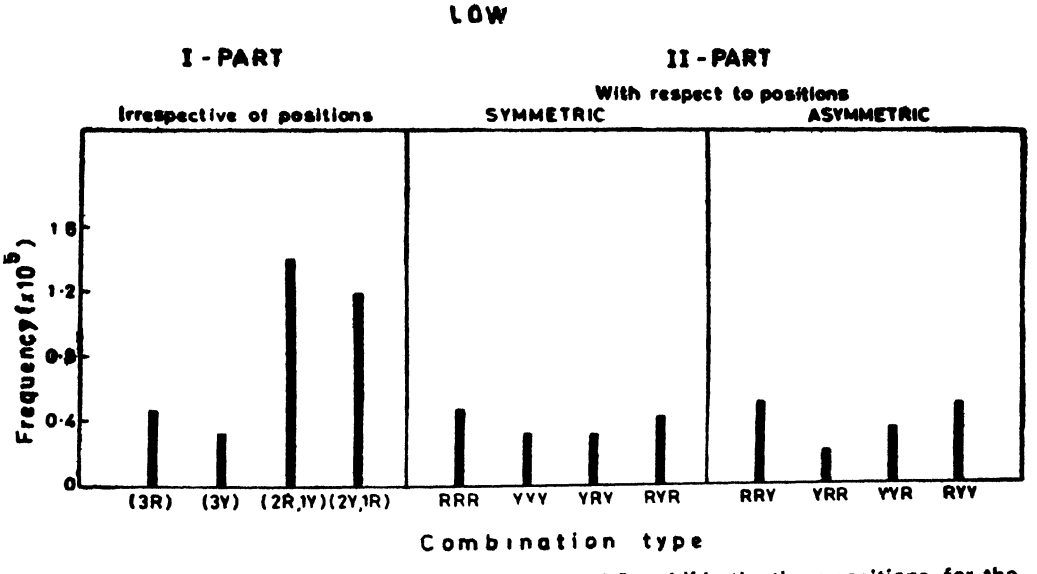

Figure 3. The frequency of occurrence of *R* and *Y* in the three positions for the LOW is reprented by histograms in II" part and their occurrences interms of symmetric and symmetric modes are represented in II" part.

With respect to symmetric mode, GLO and HIG behave in the same way but LOW shows some deviations. It is clear from the Table 3 that RRR, YYY, RYR and YRY is the order of usage for GLO and HIG. The group LOW alone has small deviations with second and fourth preferences being interchanged. Similarly in

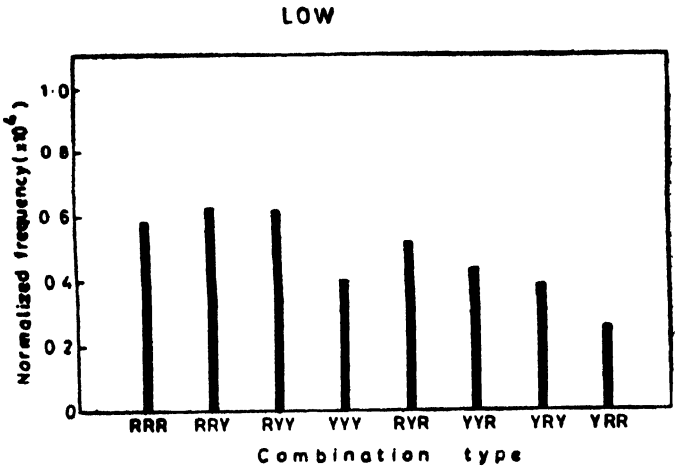

Figure 3a. The normalized frequency of occurrence of all the combination types in LOW.

Table 2. This table gives the frequencies of the various possible combination in GLO, HIG and LOW.

Combination type	GLD (×10⁵)	HIG (×10⁵)	LOW (×10⁵)
RRR (3R)	1.29089 (32.31)	0.82206 (35.25)	0.46884 (28.91)
YYY (3Y)	0.91233 (22.83)	0.59179 (25.37)	0.32054 (19.27)
RRY (2R1Y)	2.63252 (21.96)	1.18366 (16.92)	1.44886 (29.03)
YYR (2Y1R)	2.74450 (22.89)	1.57135 (22.46)	1.17314 (23.51)
	Symmetric mode		
RRR	1.29089 (33.92)	0.82206 (36.01)	0.46884 (30.79)
YYY	0.91233 (23.98)	0.59179 (25.92)	0.32054 (21.06)
YRY	0.72129 (18.96)	0.40852 (17.89)	0.31276 (20.55)
RYR	0.88072 (23.15)	0.46059 (20.18)	0.42012 (27.59)
	Asymmetric mode		
RRY	1.21940 (32.30)	0.50809 (32.46)	0.71131 (32.09)
YRR	0.53239 (14.10)	0.21497 (14.48)	0.31742 (13.58)
YYR	0.81234 (21.52)	0.35867 (20.70)	0.45366 (22.65)
RYY	1.21086 (32.08)	0.50169 (32.36)	0.70917 (31.68)

*The quantities in paranthesis represent the percentage of the normalized frequencies.

asymmetric mode, the order of usage is RRY, RYY, YYR and YRR for all the three cases.

Now irrespective of symmetric/asymmetric modes, the order of usage is RRR, RRY, RYY, YYY, RYR, YYR, YRY and YRR (for GLO and HIG) (Figures 1a and 2a) and RRY, RYY, RRR, RYR, YYR, YYY, YRY, YRR (for LOW) (Figure 3a). By and

Table 3. This table gives the normalized frequencies in all the combination types.

Combination type	GLO ($\times 10^4$)	HIG ($\times 10^4$)	LOW ($\times 10^4$)
RRR	1.6136	1.0275	0.5861
RRY	1.5242	0.8891	0.6351
RYY	1.5135	0.8865	0.6271
YYY	1.1404	0.7397	0.4007
RYR	1.1009	0.5757	0.5252
YYR	1.0154	0.5671	0.4483
YRY	0.9016	0.5107	0.3910
YRR	0.6655	0.3968	0.2687

large, it is clear that more R's (i.e., A, G) are used. As terminating codons UAA, UAG and UGA come under the pattern YRR, they are found to be least used. Similarly RRR, RRY are more used as amino acids like thr, gly, lys, asn, glu, asp are more used in proteins (Srinivasan 1978, Rajan and Srinivasan 1976).

4. Conclusions

From the above analysis, it is clear that among the 64 codons that code for various amino acids, the codons in the form RRR (in GLO and HIG), RRY (in LOW) are used maximally and similarly, the codons of the form YRR is least used in all the groups.

References

Aota S, Gojobori T, Ishibashi F, Maruyama T and Ikemure T 1988 *Nucl. Acids Res. Sequences Supl.* vol 16 r391 Table-2 623

Berger E M 1978 *J. Mol. Evol.* **10** 319

Clary D O and Wolstenhome D R 1985 *J. Mol. Evol.* **22** 252

Fitch W M 1976 *Science* **194** 1173

Hanai R and Wada A 1988 *J. Mol. Evol.* **27** 321

Haran T E, Shakked Z, Wang A H-J and Rich A 1987 *J. Biomol. Struct. Dynamics* **5** 199

Jaroslav Kypr and Jan Mrazck 1987 *Nature* **327** 20

Jukes T H 1978 *J. Mol. Evol.* **11** 121

Rajan S S and Srinivasan R 1976 *Curr. Sci.* **45** 859

Srinivasan R 1978 *Indian J. Biochem. Biophys.* **15** 75

Wachters Hauser G 1988 *Proc. Natl. Acad. Sci.* **85** 1134